Jaskaran Singh

thejaskaranbhatia@gmail.com · 437-986-0064

thejaskaranbhatia.com · linkedin.com/in/jaskaran-bhatia · github.com/jaskaranbhatia · medium.com/@jaskaranbhatia

Edge AI · LLM Inferencing · RAG · Agentic AI · Computer Vision · NLP · AWS Deployment & Architecture

Experience

J-Squared Technologies May 2023 - Present

Senior Machine Learning Engineer

- Developed an agentic workflow combining diffusion models, foundation vision models (Grounding DINO, SAM), and LLM decision agents (Llama3.2-3B and Phi3-3.8B) for autonomous generation and annotation. This has till date produced 100K+ annotated samples and reduced manual labeling effort by 90%.
- Built a RAG system for autonomous cross-team sprint summarization, integrating Notion via MCP for source ingestion, Neo4j for cross-team knowledge graph reasoning, and Ollama serving GPT-OSS-20B (native MXFP4) for local quantized LLM inference. Cut reporting time from 2+ hours to under 30 minutes.
- Development of FalconVeo, an Agentic Video RAG. The tool is built on an on-device video search system (CLIP + MS-TEMBA + Qwen3.5-VL) for privacy-preserving industrial clip retrieval. I will be presenting this project at CANSEC 2026, which is one of the largest military and defence conferences in the world.
- Vivado-MCP - Developed an AI FPGA tool for coding assistance. Shipped an internal MCP server exposing 67 Xilinx Vivado TCL (AMD's tool for FPGA Development) tools to local LLMs via Ollama, powering AI-driven HDL / testbench / constraint generation and timing closure workflows for the FPGA team, fully on-device.
- GPU Kernel & Model Optimization - Drove end-to-end performance optimization across detection, segmentation, re-identification, and pose estimation models via quantization (PTQ+QAT), pruning, knowledge distillation, and custom CUDA / TensorRT kernels. Met sub-5ms latency budgets on Hailo-8 and NVIDIA Jetson for edge deployments across clients including Retail Mining, Manufacturing, and Defence sectors.
- Built a lock-free shared-memory ring buffer in C++ as the IPC backbone for multi-process vision inference on NVIDIA Jetson AGX, supporting 4× concurrent stream throughput with end-to-end pipeline latency under 25ms.
- Built a memory optimized LLM inference service in Rust using Candle and Actix-Web, then benchmarked head-to-head against Python baselines (HuggingFace, vLLM, TensorRT-LLM). Python based engines tied with rust inference engine on P50/P99 latency, but did observe 2x faster cold starts and a lower memory footprint.

JP Morgan Chase Jan 2022 - Aug 2022

Software Engineer

- Migrated customer-facing microservices for the CIB department from Angular to React with Hooks and Redux. A/B tested rollout drove +10% user engagement and faster page-load metrics, earning team recognition award.
- Designed and shipped REST APIs in .NET Core within JPMC's microservices architecture, supporting production financial transaction workflows. Added enhanced security to APIs by adding Rate Limiting, URL Escaping, CORS, and CSRF Protection.

Scaler Academy July 2021 - Dec 2021

Software Development Engineer Intern

- Rebuilt the Referral Dashboard, newsletter dashboard, and embeddable marketing widgets in React.js. Redesigned referral flow, drove 2× referral conversion rate.
- Built REST APIs and ActiveRecord models in Ruby on Rails for referral tracking, newsletter delivery, and growth analytics workflows. Authored SQL schema migrations and CI/CD via Jenkins on AWS.

Skills

- **AI / ML:** Machine Learning, Deep Learning, Data Science, NLP, Computer Vision, PyTorch, TensorFlow, LLMs, Fine-tuning LLMs, RAG, LangChain, LangGraph, HuggingFace, CUDA, ONNX, TensorRT, OpenCV
- **Software & Cloud:** AWS, Flask, FastAPI, React.js, Ruby on Rails, Node.js, Actix.rs, C++, Swift, SQL, PySpark, Hadoop, Docker, Git, Jenkins, CI/CD
- **Programming Languages:** Python, C/C++, JavaScript, Rust, Swift

Education

University of Toronto (Canada) Sep 2022 - Dec 2023
MSc in Applied Computing (MScAC) — Computer Science GPA: 4.0 / 4.0
Courses (A+ in all): Intro. to Machine Learning, Computational Imaging, Deep Learning & Neural Networks, NLP

Thapar Institute of Engineering & Technology, Patiala (India) July 2018 - June 2022
Bachelor of Engineering in Computer Engineering (B.E.) GPA: 9.55 / 10
AI Courses: Machine Learning, NLP, CV, IoT, Data Science, Building Innovative Systems, Probability & Statistics

Publication

- Singh, J., Patel, T., & Dankar, A. (2025). *An End-to-End Pipeline for Medical Image Enhancement Using GANs Architecture.* International Journal of Emerging Science and Engineering, Vol. 13, Issue 3, pp. 32–39. Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). [DOI] [GitHub]

Projects

- **Performance Analysis of LLMs for Medical Text Summarization** [Preprint] [GitHub] – Fine-tuned and benchmarked GPT-3, GPT-4, T5, BART, and Pegasus on medical summarization, comparing ROUGE / BERTScore quality and inference cost.
- **Empirical Study of Supervised & Active Learning for Classifying Papers** [Preprint] [GitHub] – Combined conventional ML with active learning and semi-supervised methods to classify mixed labeled / unlabeled research papers efficiently.
- **Fragivo.com** [Live] – AI fragrance platform on AWS using OAuth, LLMs, and Google Search-powered vision / text analysis, with advanced prompt engineering and a recommendation system for intelligent suggestions.
- **Disaster Response Message Classifier** [GitHub] – Data engineering and ML pipeline for real-time classification of disaster messages, demonstrating scalable, production-grade NLP pipeline design.

Certifications & Achievements

- **AWS Certified:** Cloud Practitioner, AI Practitioner, Machine Learning Engineer.
- **Coursera:** Deep Learning using PyTorch (IBM), AI in Marketing & Finance (University of Pennsylvania), NLP (DeepLearning.AI), Algorithmic Toolbox (UC San Diego), and more.
- **Udacity Nanodegrees:** Machine Learning Engineer, Data Scientist, Machine Learning using TensorFlow.
- **Hackathons Won:** Code For Good (JP Morgan), Smart India Hackathon (Runner-Up), Startup Punjab Hackathon.